432 Class 02

https://thomaselove.github.io/432-2025/

2025-01-16

## Today’s Agenda

1. Comparing Means
2. Comparing Rates
3. Fitting Linear Models
4. Setting Up [Lab 1](https://thomaselove.github.io/432-2025/lab1.html), due 2025-01-22 at Noon.

[Course Notes](https://thomaselove.github.io/432-notes/): most relevant material in Chapters 1-5.

## Today’s R Setup

knitr::opts\_chunk$set(comment = NA)  
  
library(janitor) # for tabyl, clean\_names, and other things  
library(naniar) # deal with missing values  
library(broom) # for tidy, glance and augment  
library(car) # for boxCox and vif  
library(Epi) # for twoby2  
library(GGally) # for ggpairs  
library(knitr) # for kable to neaten tables  
library(kableExtra) # to adjust font sizes in kables  
library(MKinfer) # for boot.t.test  
library(mosaic) # for favstats  
library(patchwork) # for combining ggplots  
library(vcd) # for mosaic (plot) and assoc (plot)  
library(easystats) # adds in lots of tools from easystats ecosystem  
library(tidyverse) # for all kinds of things, always load last  
  
theme\_set(theme\_bw()) # another option I like is theme\_lucid()

# NHANES 1982 Example (see Course Notes: Chapters 1-5 for a very similar example)

## Loading the nh1982 R data set

Available at [our 432-data page](https://github.com/THOMASELOVE/432-data)

nh1982 <- read\_rds("c02/data/nh1982.Rds")  
  
nh1982

# A tibble: 1,982 × 9  
 SEQN age educ sbp1 sbp2 sbp3 sroh hospital mentalh  
 <chr> <dbl> <fct> <dbl> <dbl> <dbl> <fct> <fct> <fct>   
 1 109266 29 College Grad 99 99 99 Good No No   
 2 109273 36 Some College/AA 116 110 115 Good No No   
 3 109291 42 College Grad 107 111 107 Fair Yes No   
 4 109297 30 Some College/AA 105 105 102 Very Good No No   
 5 109315 30 Some College/AA 118 123 125 Good No No   
 6 109317 28 Some College/AA 110 110 110 Very Good No No   
 7 109332 33 9th - 11th Grade 110 105 108 Excellent No Yes   
 8 109333 41 Some College/AA 106 107 113 Excellent No No   
 9 109336 35 College Grad 162 148 163 Good No No   
10 109373 30 Some College/AA 111 111 113 Poor No No   
# ℹ 1,972 more rows

## 2017 - March 2020 NHANES Data

1982 NHANES subjects ages 26-42 with complete data on these 9 variables:

| Variable | Source | Description |
| --- | --- | --- |
| SEQN | P-DEMO | Subject ID: Link (also in BPXO and HUQ) |
| age | P\_DEMO | RIDAGEYR (restricted to ages 26-42 here) |
| educ | P\_DEMO | DMDEDUC2 (five-category factor) |
| sbp1 | BPXO | BPXOSY1 = 1st Systolic BP reading, in mm Hg |
| sbp2 | BPXO | BPXOSY2 = 2nd Systolic BP reading |
| sbp3 | BPXO | BPXOSY3 = 3rd Systolic BP reading |
| sroh | HUQ | HUQ010 = five-categories E, VG, G, F, P |
| hospital | HUQ | HUQ071 = Yes or No |
| mentalh | HUQ | HUQ090 = Yes or No |

## Variable Descriptions (without SEQN)

| Variable | Description (n = 1982) |
| --- | --- |
| SEQN | Subject identification code from NHANES |
| age | Age in years (range 26-42, mean = 34) |
| educ | Educational Attainment in five categories (see next slide) |
| sbp1 | Systolic Blood Pressure (1st reading) |
| sbp2 | Systolic Blood Pressure (2nd reading) |
| sbp3 | Systolic Blood Pressure (3rd reading) |
| sroh | Self-reported Overall Health: five categories (see next slide) |
| hospital | Yes if hospitalized in last 12m, else No (8% Yes) |
| mentalh | Yes if saw a mental health professional in last 12m, else No (12% Yes) |

## SROH and Educational Attainment

nh1982 |> tabyl(sroh) |> adorn\_pct\_formatting()

sroh n percent  
 Excellent 294 14.8%  
 Very Good 598 30.2%  
 Good 728 36.7%  
 Fair 321 16.2%  
 Poor 41 2.1%

nh1982 |> tabyl(educ) |> adorn\_pct\_formatting()

educ n percent  
 Less than 9th Grade 90 4.5%  
 9th - 11th Grade 209 10.5%  
 High School Grad 418 21.1%  
 Some College/AA 677 34.2%  
 College Grad 588 29.7%

## Adding mean\_sbp to the data

nh1982 <- nh1982 |>  
 mutate(mean\_sbp = (sbp1 + sbp2 + sbp3)/3)  
  
names(nh1982)

[1] "SEQN" "age" "educ" "sbp1" "sbp2" "sbp3"   
 [7] "sroh" "hospital" "mentalh" "mean\_sbp"

nh1982 |> select(mean\_sbp) |> summary()

mean\_sbp   
 Min. : 76.33   
 1st Qu.:106.33   
 Median :114.67   
 Mean :116.06   
 3rd Qu.:124.00   
 Max. :209.33

favstats(nh1982$mean\_sbp)

min Q1 median Q3 max mean sd n missing  
 76.33333 106.3333 114.6667 124 209.3333 116.0557 14.35087 1982 0

## data\_codebook() results

data\_codebook(nh1982)

nh1982 (1982 rows and 10 variables, 10 shown)  
  
ID | Name | Type | Missings | Values | N  
---+----------+-------------+----------+---------------------+-------------  
1 | SEQN | character | 0 (0.0%) | 109266 | 1 ( 0.1%)  
 | | | | 109273 | 1 ( 0.1%)  
 | | | | 109291 | 1 ( 0.1%)  
 | | | | 109297 | 1 ( 0.1%)  
 | | | | 109315 | 1 ( 0.1%)  
 | | | | 109317 | 1 ( 0.1%)  
 | | | | 109332 | 1 ( 0.1%)  
 | | | | 109333 | 1 ( 0.1%)  
 | | | | 109336 | 1 ( 0.1%)  
 | | | | 109373 | 1 ( 0.1%)  
 | | | | (...) |   
---+----------+-------------+----------+---------------------+-------------  
2 | age | numeric | 0 (0.0%) | [26, 42] | 1982  
---+----------+-------------+----------+---------------------+-------------  
3 | educ | categorical | 0 (0.0%) | Less than 9th Grade | 90 ( 4.5%)  
 | | | | 9th - 11th Grade | 209 (10.5%)  
 | | | | High School Grad | 418 (21.1%)  
 | | | | Some College/AA | 677 (34.2%)  
 | | | | College Grad | 588 (29.7%)  
---+----------+-------------+----------+---------------------+-------------  
4 | sbp1 | numeric | 0 (0.0%) | [76, 205] | 1982  
---+----------+-------------+----------+---------------------+-------------  
5 | sbp2 | numeric | 0 (0.0%) | [69, 219] | 1982  
---+----------+-------------+----------+---------------------+-------------  
6 | sbp3 | numeric | 0 (0.0%) | [60, 204] | 1982  
---+----------+-------------+----------+---------------------+-------------  
7 | sroh | categorical | 0 (0.0%) | Excellent | 294 (14.8%)  
 | | | | Very Good | 598 (30.2%)  
 | | | | Good | 728 (36.7%)  
 | | | | Fair | 321 (16.2%)  
 | | | | Poor | 41 ( 2.1%)  
---+----------+-------------+----------+---------------------+-------------  
8 | hospital | categorical | 0 (0.0%) | Yes | 159 ( 8.0%)  
 | | | | No | 1823 (92.0%)  
---+----------+-------------+----------+---------------------+-------------  
9 | mentalh | categorical | 0 (0.0%) | Yes | 247 (12.5%)  
 | | | | No | 1735 (87.5%)  
---+----------+-------------+----------+---------------------+-------------  
10 | mean\_sbp | numeric | 0 (0.0%) | [76.33, 209.33] | 1982  
---------------------------------------------------------------------------

# Comparing Means ([Course Notes Chapter 3](https://thomaselove.github.io/432-notes/431review1.html))

## Paired or Independent Samples?

* In Analysis 1, we will compare the means of SBP1 and SBP2 for our 1982 participants.
* In Analysis 2, we will compare the mean of SBP3 for our 159 participants who were hospitalized to the mean of SBP3 for our 1823 participants who were not hospitalized.

Which of these analyses uses paired samples, and why?

## Paired Samples Analysis

nh1982 <- nh1982 |> mutate(SBP\_diff = sbp1 - sbp2)  
  
favstats(~ SBP\_diff, data = nh1982)

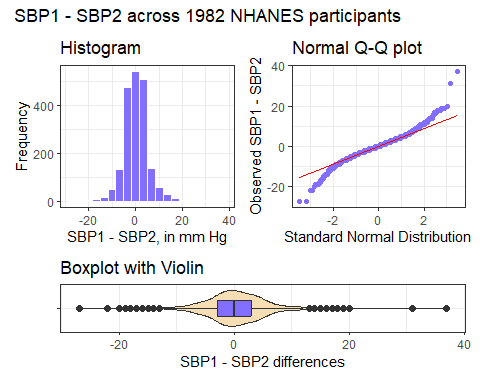
min Q1 median Q3 max mean sd n missing  
 -27 -3 0 3 37 0.2482341 5.279749 1982 0

Let’s build a set of plots to describe the distribution of SBP\_diff:

* A histogram
* A box-and-whisker plot with violins
* A normal Q-Q plot

## Paired SBP Differences

p1 <- ggplot(nh1982, aes(x = SBP\_diff)) +  
 geom\_histogram(bins = 20, col = "white", fill = "slateblue1") +  
 labs(title = "Histogram", x = "SBP1 - SBP2, in mm Hg", y = "Frequency")  
  
p2 <- ggplot(nh1982, aes(sample = SBP\_diff)) +  
 geom\_qq(col = "slateblue1") + geom\_qq\_line(col = "red") +  
 labs(title = "Normal Q-Q plot", x = "Standard Normal Distribution",  
 y = "Observed SBP1 - SBP2")  
  
p3 <- ggplot(nh1982, aes(x = SBP\_diff, y = "")) +  
 geom\_violin(fill = "wheat") +   
 geom\_boxplot(width = 0.4, fill = "slateblue1", outlier.size = 2) +   
 labs(title = "Boxplot with Violin", x = "SBP1 - SBP2 differences", y = "")  
  
(p1 + p2) / p3 + plot\_layout(heights = c(3,1)) +   
 plot\_annotation(title = "SBP1 - SBP2 across 1982 NHANES participants")



## Comparing Paired Samples

Want a 95% confidence interval for the true mean of the paired SBP1 - SBP2 differences:

* t-based approach (linear model) assumes Normality
* Wilcoxon signed rank approach doesn’t assume Normality but makes inferences about the pseudo-median
* bootstrap doesn’t assume Normality, and describes mean

set.seed(20250116)  
boot.t.test(nh1982$SBP\_diff, conf.level = 0.95, boot = TRUE, R = 999)

Results on the next two slides…

## Bootstrap 95% CI

* Estimate mean of (SBP1 - SBP2) for population based on sampled 1982 NHANES participants.
  + Sample mean SBP1 - SBP2 difference = 0.248
  + 95% CI from bootstrap: (0.020, 0.496)
  + 95% CI from t-based approach: (0.016, 0.481)
* boot.t.test() from MKinfer package results on next slide

## boot.t.test() results

Bootstrap One Sample t-test  
  
data: nh1982$SBP\_diff  
number of bootstrap samples: 999  
bootstrap p-value = 0.05205   
bootstrap mean of x (SE) = 0.2507699 (0.1185597)   
95 percent bootstrap percentile confidence interval:  
 0.01967709 0.49603935  
  
Results without bootstrap:  
t = 2.0931, df = 1981, p-value = 0.03646  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 0.01565269 0.48081552  
sample estimates:  
mean of x   
0.2482341

## Interpreting the Bootstrap CI

* The confidence interval reflects imprecision in the population estimate, based only on assuming that the participants are selected at random from the population of interest.
* When we generalize beyond study participants to the population they were selected at random from, then our data are compatible (at the 95% confidence level) with population means of SBP1 - SBP2 between 0.020 and 0.496, depending on the assumptions of our bootstrap procedure being correct.

## Comparing sbp3 by hospital: Independent Samples

favstats(sbp3 ~ hospital, data = nh1982)

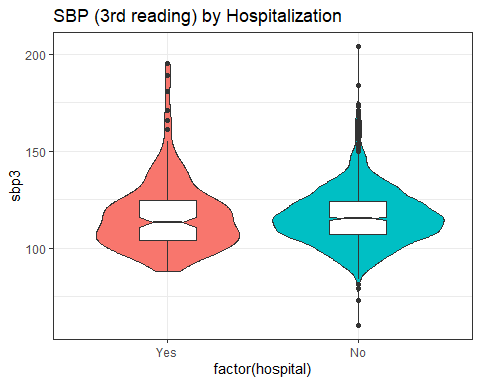
hospital min Q1 median Q3 max mean sd n missing  
1 Yes 88 104 113 124.5 195 116.7107 18.50260 159 0  
2 No 60 107 115 124.0 204 116.1130 14.50532 1823 0

* Our sample yields a point estimate for the “Hospitalized” - “Not Hospitalized” difference in means of 0.60 mm Hg.

Let’s draw a picture that lets us compare SBP3 values across the two groups.

## Comparison Boxplot, with Violins

ggplot(nh1982, aes(x = factor(hospital), y = sbp3)) +  
 geom\_violin(aes(fill = factor(hospital))) +  
 geom\_boxplot(width = 0.3, notch = TRUE) +  
 guides(fill = "none") +  
 labs(title = "SBP (3rd reading) by Hospitalization")



## Independent Samples: Comparing Means

Want a 90% confidence interval for the difference in means of SBP3 for those hospitalized - those not.

* Pooled t-based approach (equivalent to linear model) assumes Normality and equal population variances
* Welch t-based approach assumes Normality only
* bootstrap assumes neither

Suppose we’re willing to assume both Normality and equal population variances…

## Pooled t test via linear model

fit2 <- lm(sbp3 ~ hospital, data = nh1982)  
  
tidy(fit2, conf.int = TRUE, conf.level = 0.95)

# A tibble: 2 × 7  
 term estimate std.error statistic p.value conf.low conf.high  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 117. 1.18 99.0 0 114. 119.   
2 hospitalNo -0.598 1.23 -0.486 0.627 -3.01 1.81

glance(fit2) |> select(r.squared, sigma)

# A tibble: 1 × 2  
 r.squared sigma  
 <dbl> <dbl>  
1 0.000119 14.9

## Or, if you prefer…

model\_parameters(fit2, ci = 0.95) |> print\_md(digits = 3)

| Parameter | Coefficient | SE | 95% CI | t(1980) | p |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 116.711 | 1.179 | (114.399, 119.022) | 99.010 | < .001 |
| hospital (No) | -0.598 | 1.229 | (-3.008, 1.813) | -0.486 | 0.627 |

model\_performance(fit2) |> print\_md(digits = 2)

Indices of model performance

| AIC | AICc | BIC | R2 | R2 (adj.) | RMSE | Sigma |
| --- | --- | --- | --- | --- | --- | --- |
| 16327.23 | 16327.24 | 16344.00 | 1.19e-04 | -3.86e-04 | 14.86 | 14.86 |

## Or, if you prefer…

summary(fit2)

Call:  
lm(formula = sbp3 ~ hospital, data = nh1982)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-56.113 -10.012 -1.113 7.887 87.887   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 116.7107 1.1788 99.010 <2e-16 \*\*\*  
hospitalNo -0.5977 1.2291 -0.486 0.627   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 14.86 on 1980 degrees of freedom  
Multiple R-squared: 0.0001194, Adjusted R-squared: -0.0003856   
F-statistic: 0.2365 on 1 and 1980 DF, p-value: 0.6268

confint(fit2, level = 0.95)

2.5 % 97.5 %  
(Intercept) 114.398918 119.022466  
hospitalNo -3.008173 1.812791

## Interpreting the Results

* Our sample yields a point estimate for the “Hospitalized” - “Not Hospitalized” difference in means of 0.60 mm Hg, with a 95% confidence interval of (-1.8, 3.0) mm Hg.
* When we generalize beyond study participants to the population they were selected at random from, then our data are compatible (at the 95% confidence level) with a population mean difference (hospitalized - not hospitalized) in SBP3 values between -1.8 mm Hg and 3.0 mm Hg, depending on the assumptions of our linear model being correct.

# Comparing Rates (see [Course Notes, Chapter 4](https://thomaselove.github.io/432-notes/431review2.html))

## A Two-by-Two Contingency Table

nh1982 |> tabyl(mentalh, hospital) |>   
 adorn\_totals(where = c("row", "col")) |>  
 adorn\_title()

hospital   
 mentalh Yes No Total  
 Yes 37 210 247  
 No 122 1613 1735  
 Total 159 1823 1982

## Standard Epidemiological Format

nh1982 <- nh1982 |>   
 mutate(mentalh\_f = fct\_recode(factor(mentalh),   
 "Saw MHP" = "Yes", "No MHP" = "No"),  
 mentalh\_f = fct\_relevel(mentalh\_f,   
 "Saw MHP", "No MHP"),  
 hospital\_f = fct\_recode(factor(hospital),   
 "Hosp." = "Yes", "No Hosp." = "No"),  
 hospital\_f = fct\_relevel(hospital\_f,   
 "Hosp.", "No Hosp."))  
  
nh1982 |> tabyl(mentalh\_f, hospital\_f)

mentalh\_f Hosp. No Hosp.  
 Saw MHP 37 210  
 No MHP 122 1613

## Two by Two Table Analysis

twoby2(nh1982$mentalh\_f, nh1982$hospital\_f, conf.level = 0.90)

2 by 2 table analysis:   
------------------------------------------------------   
Outcome : Hosp.   
Comparing : Saw MHP vs. No MHP   
  
 Hosp. No Hosp. P(Hosp.) 90% conf. interval  
Saw MHP 37 210 0.1498 0.1161 0.1911  
No MHP 122 1613 0.0703 0.0609 0.0811  
  
 90% conf. interval  
 Relative Risk: 2.1303 1.5977 2.8405  
 Sample Odds Ratio: 2.3295 1.6723 3.2449  
Conditional MLE Odds Ratio: 2.3282 1.6287 3.2894  
 Probability difference: 0.0795 0.0442 0.1217  
  
 Exact P-value: 0.0001   
 Asymptotic P-value: 0.0000   
------------------------------------------------------

## A Larger Two-Way Table

What is the association of Educational Attainment with Self-Reported Overall Health?

nh1982 |> tabyl(educ, sroh) |>   
 adorn\_totals(where =c("row","col"))|> adorn\_title()

sroh   
 educ Excellent Very Good Good Fair Poor Total  
 Less than 9th Grade 10 7 36 33 4 90  
 9th - 11th Grade 21 40 81 59 8 209  
 High School Grad 50 94 168 98 8 418  
 Some College/AA 72 220 264 104 17 677  
 College Grad 141 237 179 27 4 588  
 Total 294 598 728 321 41 1982

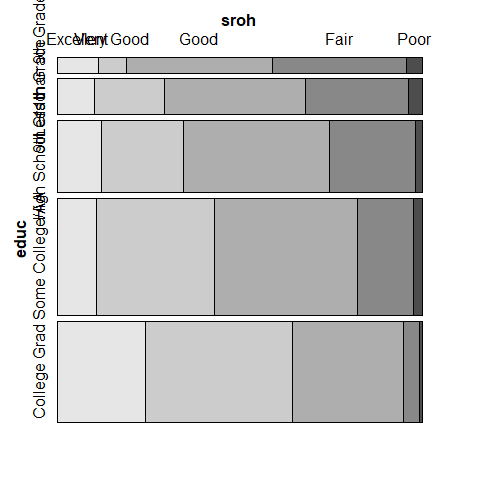
## Our 5x5 Table, showing SROH Proportions

nh1982 |> tabyl(educ, sroh) |>   
 adorn\_totals(where = c("row")) |>  
 adorn\_percentages(denominator = "row") |>   
 adorn\_pct\_formatting() |> adorn\_title()

sroh   
 educ Excellent Very Good Good Fair Poor  
 Less than 9th Grade 11.1% 7.8% 40.0% 36.7% 4.4%  
 9th - 11th Grade 10.0% 19.1% 38.8% 28.2% 3.8%  
 High School Grad 12.0% 22.5% 40.2% 23.4% 1.9%  
 Some College/AA 10.6% 32.5% 39.0% 15.4% 2.5%  
 College Grad 24.0% 40.3% 30.4% 4.6% 0.7%  
 Total 14.8% 30.2% 36.7% 16.2% 2.1%

## Mosaic Plot for our 5x5 Table

mosaic(~ educ + sroh, data = nh1982, highlighting = "sroh")



## Pearson test for our 5x5 Table

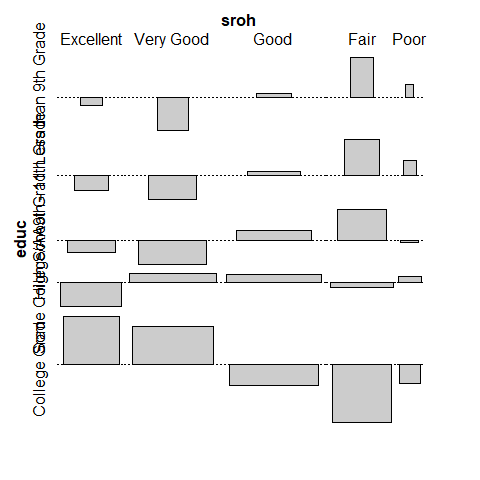
chisq.test(xtabs(~ educ + sroh, data = nh1982))

Warning in stats::chisq.test(x, y, ...): Chi-squared approximation may be  
incorrect

Pearson's Chi-squared test  
  
data: xtabs(~educ + sroh, data = nh1982)  
X-squared = 225.99, df = 16, p-value < 2.2e-16

## Association Plot for our 5x5 Table

assoc(~ educ + sroh, data = nh1982)



# Fitting Linear Models (see [Course Notes, Chapter 5](https://thomaselove.github.io/432-notes/431review3.html))

## We’ll fit two models today

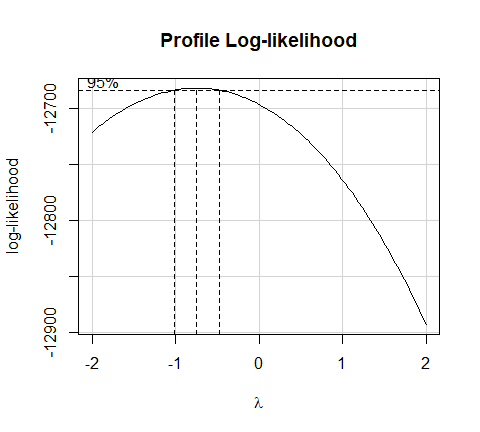
1. Predict mean SBP using Age alone.
2. Predict mean SBP (across three readings) using Age, Self-Reported Overall Health Status and Hospitalization Status.

temp\_mod1 <- lm(mean\_sbp ~ age, data = nh1982)  
temp\_mod2 <- lm(mean\_sbp ~ age + sroh + hospital,   
 data = nh1982)

I’m not doing any predictive validation today (remember we did that in Class 1), so I won’t split the sample.

## Box-Cox Plot to suggest potential outcome transformations

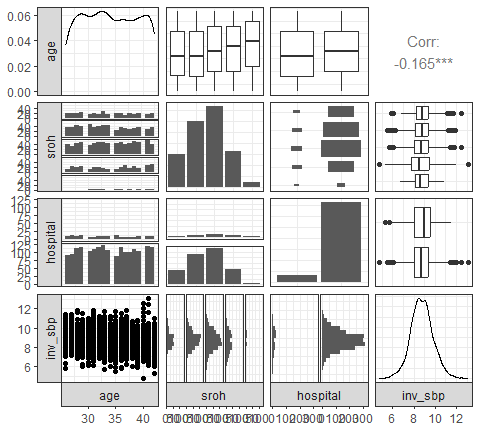
boxCox(temp\_mod2)



nh1982 <- nh1982 |> mutate(inv\_sbp = 1000/mean\_sbp)

## Scatterplot Matrix (from ggpairs())

ggpairs(nh1982, columns = c(2, 7, 8, 14), switch = "both",  
 lower=list(combo=wrap("facethist", bins=20)))



## Variance Inflation Factors

car::vif(lm(inv\_sbp ~ age + sroh + hospital, data = nh1982))

GVIF Df GVIF^(1/(2\*Df))  
age 1.008723 1 1.004352  
sroh 1.020544 4 1.002545  
hospital 1.013797 1 1.006875

## Tidied Coefficients for Model m1

m1 <- lm(inv\_sbp ~ age, data = nh1982)  
  
tidy(m1, conf.int = TRUE, conf.level = 0.9)

# A tibble: 2 × 7  
 term estimate std.error statistic p.value conf.low conf.high  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 9.93 0.161 61.5 0 9.66 10.2   
2 age -0.0349 0.00469 -7.44 1.51e-13 -0.0426 -0.0272

### Model Parameters for m1

model\_parameters(m1, ci = 0.9)

Parameter | Coefficient | SE | 90% CI | t(1980) | p  
------------------------------------------------------------------------  
(Intercept) | 9.93 | 0.16 | [ 9.66, 10.20] | 61.52 | < .001  
age | -0.03 | 4.69e-03 | [-0.04, -0.03] | -7.44 | < .001

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

## Tidied Coefficients for Model m2

m2 <- lm(inv\_sbp ~ age + sroh + hospital, data = nh1982)  
  
tidy(m2, conf.int = TRUE, conf.level = 0.9)

# A tibble: 7 × 7  
 term estimate std.error statistic p.value conf.low conf.high  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 10.0 0.185 54.3 0 9.74 10.3   
2 age -0.0338 0.00470 -7.19 9.27e-13 -0.0415 -0.0260   
3 srohVery Good -0.0552 0.0727 -0.759 4.48e- 1 -0.175 0.0644   
4 srohGood -0.110 0.0705 -1.56 1.20e- 1 -0.226 0.00627  
5 srohFair -0.265 0.0825 -3.21 1.33e- 3 -0.401 -0.129   
6 srohPoor -0.176 0.171 -1.03 3.02e- 1 -0.457 0.105   
7 hospitalNo -0.0464 0.0849 -0.546 5.85e- 1 -0.186 0.0933

## Model Parameters for m2

model\_parameters(m2, ci = 0.9)

Parameter | Coefficient | SE | 90% CI | t(1975) | p  
-----------------------------------------------------------------------------  
(Intercept) | 10.04 | 0.18 | [ 9.74, 10.34] | 54.32 | < .001  
age | -0.03 | 4.70e-03 | [-0.04, -0.03] | -7.19 | < .001  
sroh [Very Good] | -0.06 | 0.07 | [-0.17, 0.06] | -0.76 | 0.448   
sroh [Good] | -0.11 | 0.07 | [-0.23, 0.01] | -1.56 | 0.120   
sroh [Fair] | -0.27 | 0.08 | [-0.40, -0.13] | -3.21 | 0.001   
sroh [Poor] | -0.18 | 0.17 | [-0.46, 0.10] | -1.03 | 0.302   
hospital [No] | -0.05 | 0.08 | [-0.19, 0.09] | -0.55 | 0.585

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

## Compare Coefficients: m1 and m2

compare\_models(m1, m2)

Parameter | m1 | m2  
--------------------------------------------------------------  
(Intercept) | 9.93 ( 9.61, 10.25) | 10.04 ( 9.68, 10.40)  
age | -0.03 (-0.04, -0.03) | -0.03 (-0.04, -0.02)  
sroh [Very Good] | | -0.06 (-0.20, 0.09)  
sroh [Good] | | -0.11 (-0.25, 0.03)  
sroh [Fair] | | -0.27 (-0.43, -0.10)  
sroh [Poor] | | -0.18 (-0.51, 0.16)  
hospital [No] | | -0.05 (-0.21, 0.12)  
--------------------------------------------------------------  
Observations | 1982 | 1982

## Fit Summaries for Models m1 and m2

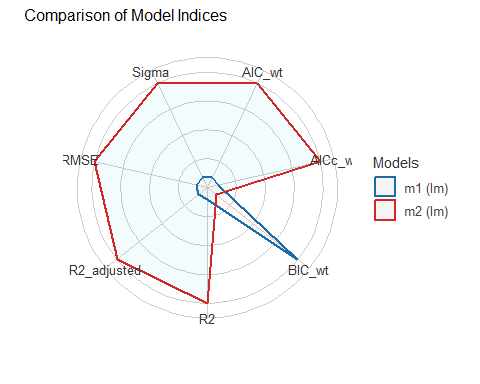
bind\_rows(glance(m1), glance(m2)) |>  
 mutate(model = c("m1", "m2")) |>   
 select(model, r2 = r.squared, adjr2 = adj.r.squared,   
 sigma, AIC, BIC, nobs, df, df.residual)

# A tibble: 2 × 9  
 model r2 adjr2 sigma AIC BIC nobs df df.residual  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <int>  
1 m1 0.0272 0.0267 1.02 5714. 5731. 1982 1 1980  
2 m2 0.0334 0.0304 1.02 5711. 5756. 1982 6 1975

Which model appears to fit the data better?

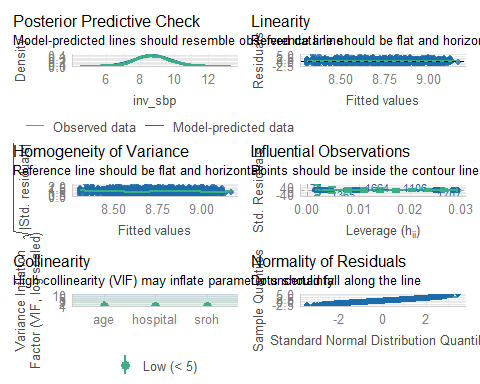
## Compare m1 to m2

plot(compare\_performance(m1, m2))



## Residual Plots for Model m2

check\_model(m2, detrend = FALSE)



## Making a Prediction in New Data

Suppose a new person is age 29, was not hospitalized, and their SROH is “Good”. What is their predicted mean systolic blood pressure?

* Our models predict 1000/mean\_sbp and augment places that prediction into .fitted.
* To invert, divide .fitted by 1000, then take the reciprocal of that result. That’s just 1000/.fitted.

## Making a Prediction in New Data

new\_person <- tibble(age = 29, sroh = "Good", hospital = "No")  
bind\_rows(augment(m1, newdata = new\_person),   
 augment(m2, newdata = new\_person)) |>  
 mutate(model = c("m1", "m2"), fit\_meansbp = 1000/.fitted) |>  
 select(model, fit\_meansbp, .fitted, age, sroh, hospital)

# A tibble: 2 × 6  
 model fit\_meansbp .fitted age sroh hospital  
 <chr> <dbl> <dbl> <dbl> <chr> <chr>   
1 m1 112. 8.92 29 Good No   
2 m2 112. 8.90 29 Good No

# Setting Up Lab 1, due 2025-01-22 at Noon

## Lab 1 Question 1

I provide some County Health Rankings data for 30 variables and 2054 counties included in the CHR 2024 report. You will filter the data down to the 88 counties in Ohio, and check for missing values.

Then you will create a visualization involving information from three different variables (from a list of 15) using R and Quarto.

There is a [Quarto template for Lab 1](https://raw.githubusercontent.com/THOMASELOVE/432-data/refs/heads/master/data/432_lab1_template.qmd), in addition to the data set.

## Lab 1 Question 2

Create a linear regression model to predict obesity as a function of food\_env, adjusting for unemployment (all of these are quantitative variables.)

1. Specify and fit the model, interpret food\_env coefficient and its confidence interval carefully.
2. Evaluate quality of model in terms of adherence to regression assumptions via check\_model().
3. Build a nice table comparing your model to a simple regression for obesity using only food\_env, then reflect on your findings.

## Coming Up…

* TA office hours begin this Friday 2025-01-17. (No hours on Monday 2025-01-20 - MLK Holiday.)
* Lab 1 due Wednesday 2015-01-22 at Noon to [Canvas](https://canvas.case.edu/)
  + Answer Sketch available 48 hours post-deadline
* Linear and Logistic Regression and the SUPPORT study